College of Information Science and Technology

Measuring the Quality of Network Visualization

Chaomei Chen
College of Information Science and Technology
Drexel University
Philadelphia, PA 19104-2875, USA
+1 215 895 6627
chaomei.chen@cis.drexel.edu

ABSTRACT
A quantitative method is developed for measuring the quality of network visualizations in terms of log-likelihood metrics resulted from Expectation Maximization (EM) clustering intrinsic and extrinsic attributes of network nodes.

Categories and Subject Descriptors
H.5.2 [**Information Systems**]: User Interfaces – *Graphical user interfaces*

General Terms
Algorithms, Measurement, Verification.

Keywords
Network visualization, quality metrics of quality, EM clustering.

1. THE RESEARCH QUESTION
Information visualization has the potential of revolutionizing the way we access scientific knowledge [1]. We introduce a quantitative approach to the evaluation of the quality of network visualizations. The quality of a visualized network is defined in terms of the probability that clusters of nodes are generated by a set of latent variables. The assumption is that if a network abstraction of the underlying data conforms to independently identified clusters in the data, then the visualized network preserves the intrinsic structure of the data.

In bibliographic networks, for example, intrinsic attributes of nodes include author info and the year of publication. Extrinsic attributes of a node include how many co-citation links connected to the node and how many shortest paths the node belongs to. Therefore, extrinsic attributes can be used to characterize topological properties of the network in question.

2. A QUANTITATIVE APPROACH
A series of networks of the same bibliographic dataset are generated by systematically altering the configuration of network modeling parameters such as the inclusion of concept terms, network scaling, and centrality metrics. A machine-learning technique Expectation Maximization (EM) clustering is used to cluster nodes based on attributes computed in each configuration. The log-likelihood metric of EM clustering quantifies the probability that the observed data is generated by a latent mixture model. The higher the log-likelihood value ***L***, the better the network modeling configuration. The network modeling, visualization, and EM clustering functions are implemented in the latest version of CiteSpace [2, 3].

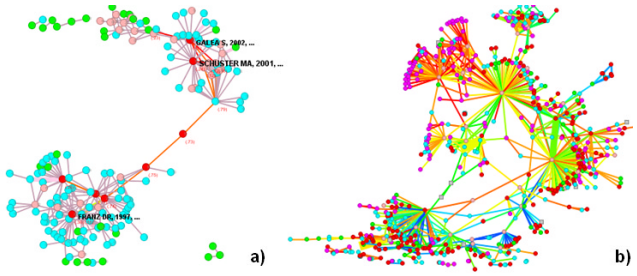

Figure 1. The quality of visualization a) is better (log likelihood: L = -11.8506) than that of b) (L = -14.5967).

3. RESULTS AND FUTURE WORK
We demonstrate the potential value of a cross validation method built on distance-based networks and model-based clustering algorithms in machine learning. The question of searching for a quality metric is transformed to a question of what the likelihood that the given data is generated by such mixture models is.

The main contribution of the work is the first step to bridge potentially useful techniques from different communities and to develop a quality metric that can be used and refined for a number of network visualization. In the context of digital library research, the provision of such metrics has practical as well as theoretical implications because it can facilitate users to select an optimal configuration of network visualization.

REFERENCES
1. Chen, C. *Information Visualization: Beyond the Horizon*. London: Springer Verlag.
2. Chen, C. Searching for intellectual turning points: Progressive Knowledge Domain Visualization. *Proc. Natl. Acad. Sci. USA*, *101* (2004), 5303-5310.
3. Chen, C. *CiteSpace*. http://cluster.cis.drexel.edu/~cchen

JCDL'05, June 7–11, 2005, Denver, Colorado, USA
ACM 1-58113-876-8/05/0006.